

Мифы VS реальность про LLM в бизнесе

СПИКЕР:

Захарищев Михаил
Директор по продажам

Назаркин Арсений
Технический директор



Компания в цифрах

KVINT



420+ Внедрено более 420 роботов

90+ Команда разработки роботов

7 лет На рынке

2000+ Более 2000 звонков в минуту

7 млн + Более 7 млн звонков в день

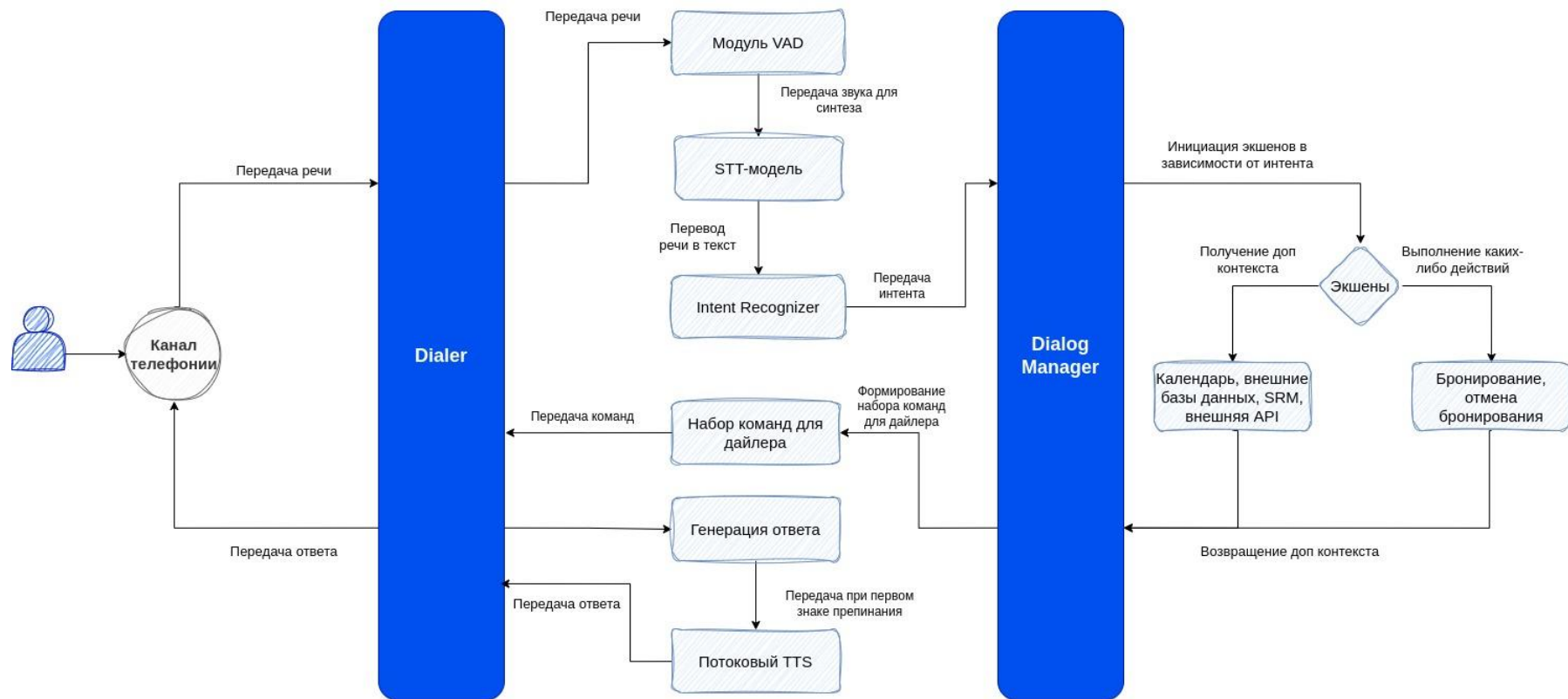
520+ млн Звонков обработали роботы КВИНТ в 2024 году

Почему только LLM недостаточно?

01 / 06



Как выглядит генеративный голосовой робот



Что делать с задержками и синтезом?

02 / 06



Скорость ответа зависит от нескольких систем

Распознавание речи
(VAD + STT)

01

Определение намерения
(NLU + LLM сервис)

02

Запросы в базу
знаний и внешние
системы

03

Генерация аудио
ответа (LLM + TTS)

04

Показатели задержек на 2025 год

**1.8 -
2.7 сек**

**Общая
длительность
генерации ответа**

0.5-0.8 сек

**Скорость
распознавания речи**

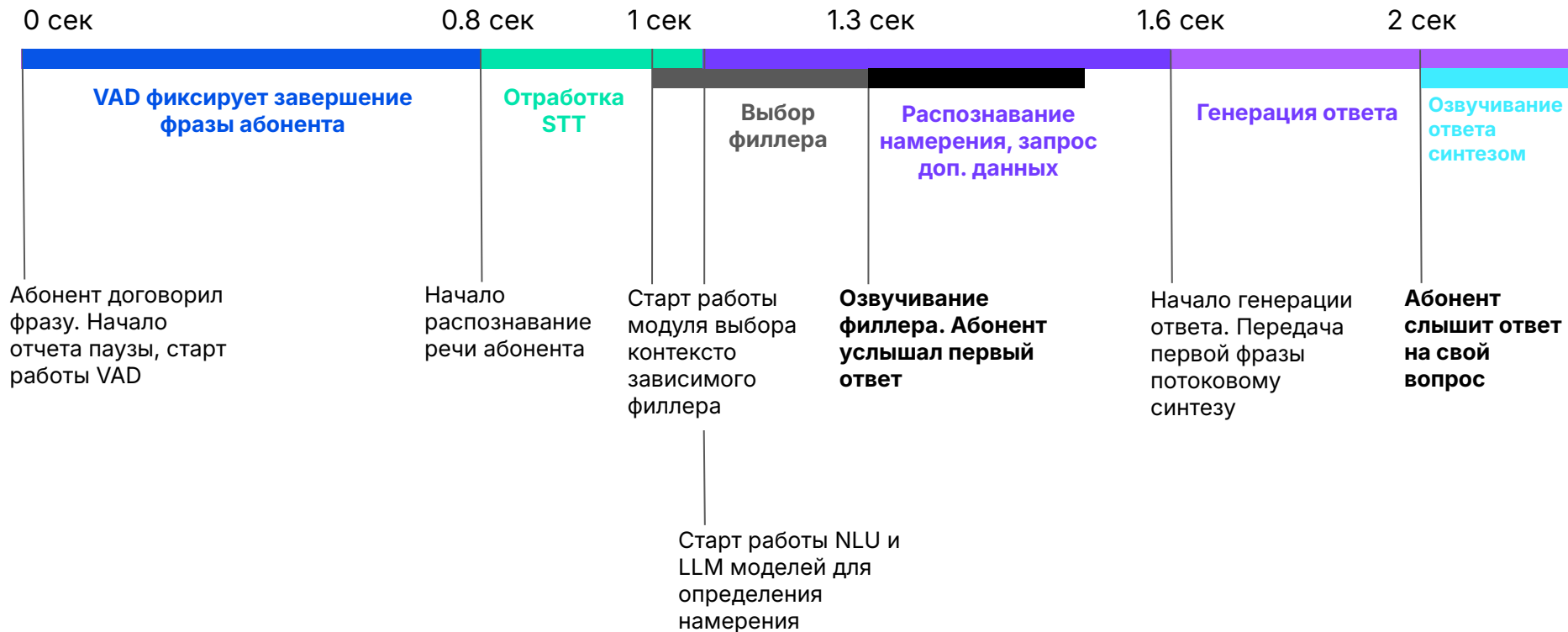
< 0.3 сек

**Определение намерения и
запрос доп данных**

1 - 1.6 сек

Генерация аудио ответа

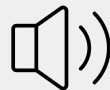
Таймлайн генерации ответа



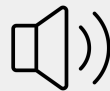


Эмоциональный синтез речи

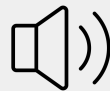
Полное копирование интонаций
референсной записи.



Запись с бесшовными
склейками



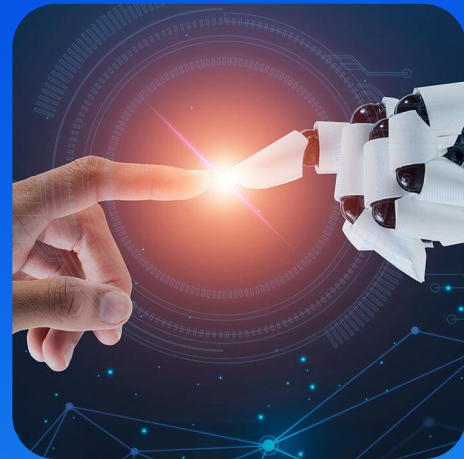
Безэмоциональный синтез



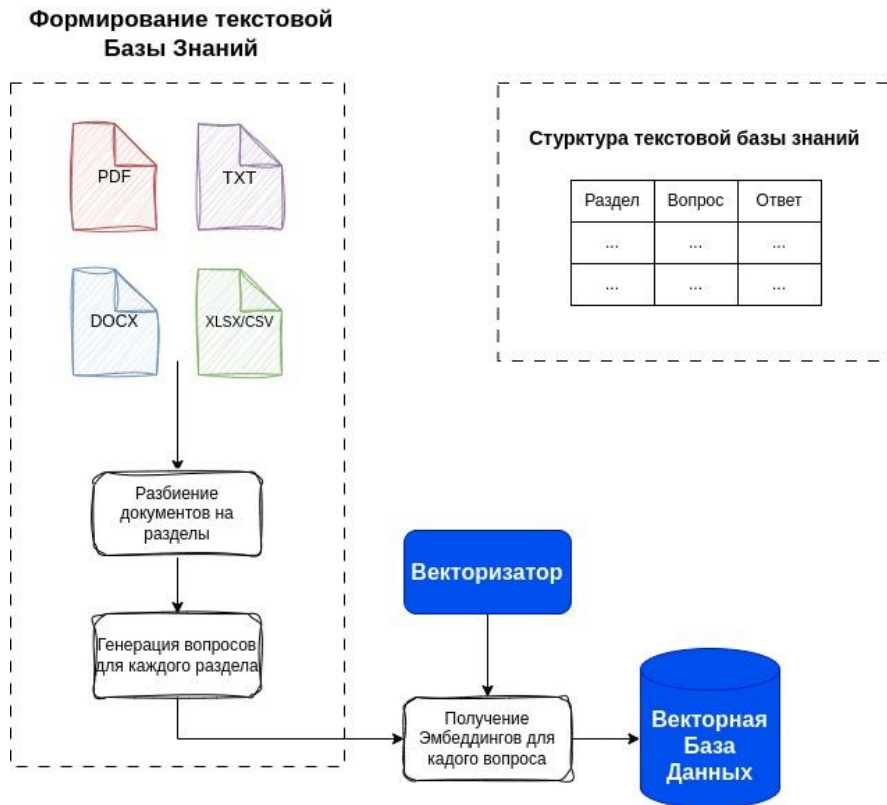
Эмоциональный синтез

Почему мы обучаем LLM сами?

03 / 06

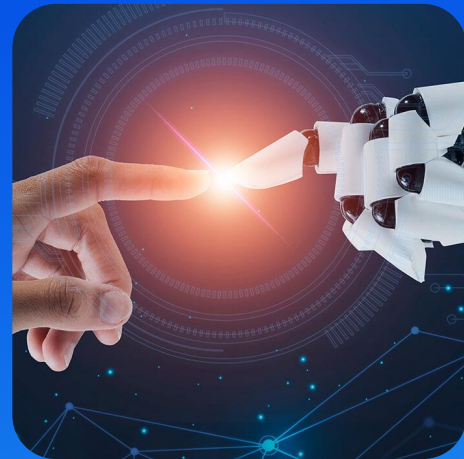


Процесс формирования базы знаний



Как контролировать качество робота?

04 / 06



Механизм контроля качества

1. Проверка работа на других текстовых ботах

Генерируем 1000+ диалогов с пред созданными персонажами и ищем ошибки с помощью отдельной LLM

2. Повторяем первый пункт, но уже в голосе

Оцениваем задержки и перебивания

3. Переходим к аналитике звонков на проде



Плохо слышит и переспрашивает



Предпочитает односложные фразы



Использует много сленга

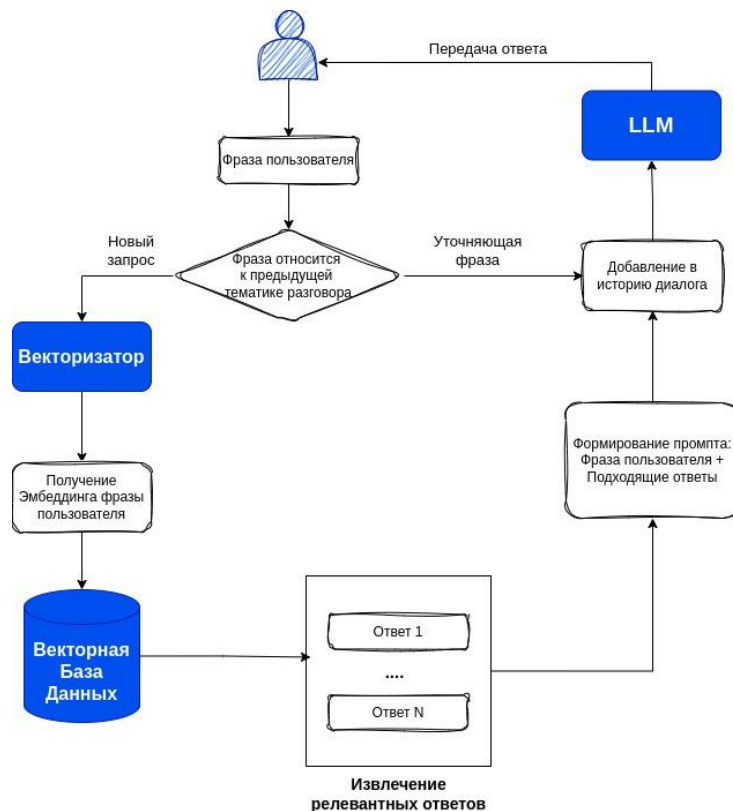
Что делать с галлюцинациями?

05 / 06



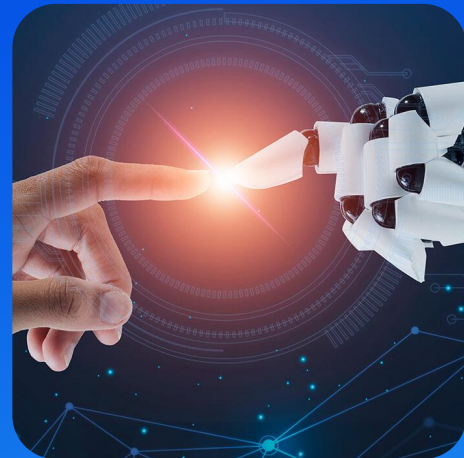
Процесс борьбы с галлюцинациями модели

1. Предварительное тестирование LLM на выдачу RAG (пустая, релевантная, нерелевантная)
2. Замеряем процент выдачи удовлетворительных ответов
3. Тюним навык с помощью обучения модели на примерах диалогов



Генеративный робот – это долго?

06 / 06



Как выглядит процесс внедрения



Контакты

Захарищев Михаил

Директор по продажам

+7 922 931 1888

m.zak@kvint.io

www.kvint.io

